Toward a Passenger-Oriented Model of Subway Performance

GARY HENDERSON, HEBA ADKINS, AND PHILIP KWONG

On-time performance measures used by transportation operating agencies typically use definitions, procedures, and report formats that represent an operational rather than passenger-oriented perspective. Although providing a useful barometer of operational effectiveness, such systems only indirectly measure the passengers' experience of service. Using a random sampling methodology to construct a computerized data base of about 50,000 morning rush hour subway trains, the subway performance model developed by New York State's Metropolitan Transportation Authority Inspector General's Office is designed to measure service as subway passengers experience it. The system focuses on actual, not scheduled, service; it measures aspects of service most meaningful to riders, in terms they can relate to, and on a scale experienced by passengers. Measuring performance according to this principle affects every aspect of research design and analysis, including the selection of measurement points, the definition of a trip and a route, the time periods used, the scale of analysis (system, route, or more detailed), and the statistics to be reported. The basic concept also entails a reconsideration of the way train cancellations, bypasses, service adjustments, extra service, and headway irregularities are treated in measuring on-time performance. Features of the methodology resolve many of these analytical issues, while presenting numerous avenues for further research and development.

Transit agencies typically produce performance measures for two distinct purposes. Some statistics are calculated for internal, operational purposes, while others are produced for public reporting. There is considerable overlap and tension between these two directions. Measures are often amalgams of the two aims. Statistics produced from the perspective of transit operations are often useful for public reporting, particularly to oversight agencies and legislative bodies. Similarly, statistics produced from the public's or passenger's perspective are also useful for internal diagnostic purposes and strategic management goals.

A conceptual approach to measuring performance is being developed by the New York State Metropolitan Transportation Authority's Inspector General's Office (MTA–IG). The MTA–IG's approach to on-time performance is contrasted with the measurement system currently used by the New York City Transit Authority (NYCTA). Because there are about 3.7 million passengers riding the New York subway system each weekday, the passengers' perspective is considered. Although statistics cannot describe the experience of every passenger, the MTA–IG model provides an analytical framework to distinguish between discrete groups of passengers when calculating on-time performance.

New York State Metropolitan Transportation Authority, Office of the Inspector General, 100 Park Ave., 14th Floor, New York, N.Y. 10017.

A simple example of the difference between the operational and passenger perspectives is the calculation of the delay that occurs when a train breaks down between stations. From the operational perspective, the delay is over once the passengers are discharged and the disabled train departs. For the riders, the delay is over when they board another train.

NYCTA performance measures at present do not meet the requirements of a passenger-oriented model. For example, they focus on the percentage of trains arriving on time, not on the percentage of passengers arriving at their destinations on time. In many cases, train arrivals are measured at remote terminal locations, not at the main stations used by passengers as destinations. On-time performance (OTP) statistics are grouped in large time intervals (e.g., a.m. rush hour or 24 hr), not the smaller intervals corresponding to a passenger's routine commute. Extra service provided to passengers is not integrated into reported OTP measures, and cancelled service and bypasses are considered delays, even when passengers do not arrive late at their destinations. Schedule adjustments made by dispatchers make trains appear to be on time, even when travel times or platform waiting times are increased. Estimates of average passenger delays are not made or reported by the NYCTA.

The performance model developed by the MTA–IG is designed to measure service as passengers see it, while preserving the capability of producing operational measures. A data base of approximately 50,000 trains, reaching their respective central business district (CBD) stations between 6:00 and 10:00 a.m., was constructed from subway records. The passenger orientation influenced every aspect of model design. To illustrate the methods discussed, 1988 data were analyzed for several subway services (a total of 5,103 trains): the No. 1 southbound (1,147 trains), the F northbound (700 trains), both the No. 4 northbound (926 trains) and southbound (907 trains), and both the No. 5 northbound (546 trains) and southbound (877 trains). All northbound No. 5 trains from Utica Avenue were treated as No. 4 trains, because they are identical services until after the CBD.

It was originally estimated that only half as many trains were needed for the reliable No. 1 line, but on one of the randomly chosen sample days, a derailment caused about half the scheduled service to be cancelled. The elimination of these data could not be justified because the passenger's perspective is inconsistent with the concept of a typical day. Passengers ride the subway on all days, so it is important to preserve the full range of experience. However, the sample size for the No. 1 was doubled to ensure a higher level of accuracy for the statistical estimates. In the final analysis, the derailment caused the annual averages to drop 2 to 4 percent.

CHOOSING WHERE TO MEASURE

Two of New York's commuter railroads measure morning rush OTP at Penn Station and Grand Central Station, which are terminals as well as the primary destinations for their peak ridership. The situation is not as convenient for the subway system, because most lines have terminals at outlying locations. Though most passengers have deboarded trains at various CBD locations, the NYCTA measures OTP some time later at outlying terminals, where trains have a smaller and altogether different ridership. For example, the NYCTA monitors F trains from Brooklyn at 179th Street in Queens, where they are scheduled to arrive 40 min after departing from the CBD point used in the MTA–IG model. This diagram depicts the F line:

T1———*I*——————*CBD*——————*T2*

T1 represents the originating terminal (Coney Island, Brooklyn), *T2* represents the final terminal (179th Street, Queens), *CBD* represents the midtown CBD point used in the MTA–IG model (West 4th Street), and *I* represents an intermediate point used in the model (Kings Highway).

From an operational point of view, the NYCTA must be concerned with the entire trip of the train, but its method combines two different trips from the passenger's point of view—one heavily loaded inbound trip to the CBD and one much more sparsely loaded outbound trip. Ideally, terminal-to-terminal trips should be measured as two distinct trips, weighted by the number of passengers on each leg of the train's trip. The MTA–IG model does not yet measure service provided for reverse commuters, but accounting for outbound trips remains a necessary future step.

In the morning rush hour, travel times to CBD locations are more relevant to passengers than travel times of trains to remote terminals. However, the travel time from originating terminals to CBD locations (from *T1* to *CBD*) does not represent the travel time of all riders, because many board trains near the CBD. For this reason, it is advisable to choose one or more intermediate control points (such as point *I* in the diagram). In this way, OTP can be calculated for two sets of passengers, those boarding between *T1* and *I* and those boarding between *I* and *CBD*.

Moreover, headways at the intermediate point are preferable to calculating platform waiting times, which are needed for the OTP analysis described later. Headways in the CBD are not as useful for measuring waiting times, because congestion occurring after many passengers have boarded can change the headway distribution significantly.

PASSENGER GROUPS, NOT TRAINS

One important feature of the MTA–IG model allows analysis of the effect of schedule adjustments on passengers, provides for a more appropriate treatment of cancellations and bypasses, and produces realistic information on delays experienced by passengers. This is the concept of passenger groups.

Each scheduled train is seen as representing a group of passengers (or two groups, when using the intermediate station *I* as a control point), and each group must be accounted for separately. When a train is cancelled, the model's computer record does not delete it; rather, it continues to record and calculate statistics for the passengers who would have been on the train. The MTA–IG model imputes a delay figure to those passengers on the basis of the arrival time of the train on which the passengers are estimated to have reached their destination. The decision rule was made that passengers arrived on the next available train. For example, if a cancelled F train's scheduled time at West 4th Street was 8:20, and the next F train arrived at 8:32, a 12-min delay was imputed to the passengers left on the platform because of the cancelled train. In this way, the gaps in service caused by cancellations can be included in calculations of the average delay.

Although passengers in New York cannot always board the next train, no empirical evidence specifies an alternative assumption. Research is now planned to develop this aspect of the model. Until more is known, the next-available-train rule will be used; it is simple, and the direction of error is known. (The severity of the delay is underestimated, especially on the most crowded lines.)

Providing a delay figure for cancelled trains allows a more accurate treatment of bypasses. Occasionally, trains are routed onto an express track to avoid congestion or a disabled train; the NYCTA categorizes bypasses as cancellations, late by definition. A portion of the line's riders were delayed, depending on the number of stops missed. However, passengers on the train may not be so delayed, because bypassing trains often arrive in the CBD on time or ahead of schedule. Moreover, directing the lead train to bypass several stops when a long delay has occurred may help restore service more quickly than having it make all stops at overcrowded platforms; the NYCTA calls this strategy a "battery run."

The MTA–IG model allocates the delay to specific trains based on the severity of the incident. For example, if five stops were bypassed by four trains, the first and last trains might be assigned the delay experienced by passengers who were left at the bypassed stops and the other two would be treated according to their actual travel time (these trains may still be late). This technique treats some of the passengers as late and accounts for the positive results of the bypassing train as well.

VARIATIONS IN OTP STATISTICS

Although a variety of reliability measures are used in the performance model, this section focuses on ways to measure OTP. The following measures of OTP will be discussed:

- Standard OTP,
- Passenger-weighted OTP,
- Operational OTP,
- Total-trip OTP, and
- Weighted total-trip OTP.

Standard OTP

Because New York City subway data are being used and the NYCTA definitions and formulas are known, its method will be used as standard. The TA defines OTP by the following equation:

$$\text{OTP} = \frac{S - L - EC - TC}{S}$$

where

S = total number of trains scheduled,
L = number of trains late by more than 5 minutes,
EC = number of trains cancelled en route, and
TC = number of trains cancelled at the terminal.

Late trains (L) are those that arrive at their destination terminal more than 5 min behind schedule. (The schedule may have been adjusted according to operating procedures designed to even out service in the event of a cancellation or delay). An en route cancellation (EC) may be a train removed from service or simply a train that bypassed one or more scheduled stops. A terminal cancellation (TC) can be caused by shortage of equipment or personnel, or simply a late arrival in one direction that causes a delay in the turnaround trip in the opposite direction. Terminal cancellations are often compensated by dispatching an extra train, but the NYCTA does not include extras in the calculation of OTP.

The MTA–IG methodology makes three significant alterations to this standard method. First, rush-hour OTP is measured at CBD locations instead of destination terminals. Second, smaller time intervals are used—half-hours for the morning rush. The NYCTA's aggregation of all trips from 7:00 to 9:00 a.m. into a single measure does not provide operations managers with sufficient information about when problems occur. If reported publicly, it would not provide passengers with useful information about the service when they ride. Third, OTP is measured for each direction of a line, whereas the NYCTA combines statistics for a single measure of each line. The differences between directions on a line are significant and are more relevant to passengers in this form.

Table 1 presents the OTP of several services calculated with the standard definition, using half-hour intervals. The OTP exhibits a striking degree of variation across the rush hour that is lost by aggregating the results into a single measure for the 7:00 to 9:00 a.m. period. The No.4 north and No.5 both north and south are particularly significant. The difference for the relatively steady F line between the 7:00 to 9:00 a.m. average and the critical 8:30 to 9:00 a.m. period is 8 percent; between the best and worst half-hour periods the difference is 25 percent. For the No.5 south, the difference between the peak half-hour and the 7:00 to 9:00 average is 35 percent; between the best and worst half-hours, the difference is 68 percent.

Table 1 also indicates the variation of OTP for different directions of the same line. The northbound No.4 is on time 36 percent from 9:00 to 9:29 and 64 percent from 9:30 to 9:59; the southbound No.4 is on time 52 percent and 92 percent for the same time periods. The No.5 service differs dramatically by direction for every time period after 7:00.

Passenger-Weighted OTP

A primary goal of a passenger-oriented system is the measurement of the percentage of passengers on time instead of the percentage of trains on time. A simple approach is to weight train OTP on the basis of aggregate passenger counts, to calculate the percentage of passengers who are late or on time. The heaviest weight would go to train trips with the highest ridership at the peak hour. This approach was tested by aggregating the OTP of several lines for the 7:00 to 9:00 a.m. period using the standard method and comparing the result to the measure reached by weighting each half-hour period for each line by passenger count. The two results were nearly identical. Thus, the standard OTP method already includes a form of passenger weight, because the frequency of trains corresponds to ridership volume.

Another form of passenger weight is to divide the trip in two with data at the intermediate station. Are riders at the more remote areas on the line more often late than those boarding closer to the CBD? Passenger counts were used to estimate the ridership for each half-hour interval before and after station *I*. Then the standard OTP was calculated for riders boarding before *I* (actual travel time from *T1* to *CBD* minus the scheduled travel time) and riders boarding after *I* (actual travel time from *I* to *CBD* minus the scheduled travel time) and weighted by the appropriate passenger loads.

When this method was applied to the F line's performance, the results were nearly identical to those of the standard method.

TABLE 1 STANDARD OTP (PERCENT)

Scheduled Arrival at CBD Location (a.m.)	F North	No.1 South	No.4 North	No.4 South	No.5 North	No.5 South
6:00-6:29	75.9	91.5	90.9	92.4	—[a]	92.4
6:30-6:59	75.0	93.8	90.9	90.1	86.4	85.2
7:00-7:29	77.8	90.8	89.8	85.5	95.4	83.6
7:30-7:59	83.3	89.9	82.1	80.2	87.1	78.0
8:00-8:29	64.8	91.1	86.4	77.9	84.8	60.7
8:30-8:59	64.3	95.3	60.5	50.6	57.3	23.8
9:00-9:29	63.0	90.5	36.3	52.2	39.8	52.2
9:30-9:59	88.6	92.2	63.6	92.0	63.5	87.5
7:00-9:00	72.0	92.0	76.5	72.3	79.9	59.3

NOTE: All northbound No.5 trains from Utica Avenue or New Lots Avenue are grouped with the northbound No.4. Excluding the sample day when a No.1 train derailed would raise the No.1 line's performance 2 to 4 percentage points in individual half-hour periods and 3 points for the 7:00-9:00 period. The CBD point is West 4th Street for the F; Times Square for the No.1; and Grand Central for the No.4 and No.5 in both directions.
[a]No northbound No.5 service before 6:30 a.m.
SOURCE: MTA–IG Analysis of 1988 NYCTA Interval Sheets.

This similarity suggests that delays on the F line occur after the intermediate station (*I*), at bottlenecks approaching the CBD, and that they affect all passengers.

Operational OTP

Under the definition of operational OTP, to be on time a train must arrive not later than the next scheduled arrival and be no more than 4 min late. This is called operational OTP because the definition of lateness varies with the frequency of service. If short headways are scheduled, then the criterion for on time should reflect whether one train arrives in the slot scheduled for another. This approach was proposed for the NYCTA in legislation introduced in 1989 by New York Assemblywoman Catherine Nolan. Table 2 indicates how operational OTP differs from the standard for two subway services.

The most obvious effects are that operational OTP is usually lower than standard OTP because it uses more stringent criteria, and that this difference grows in the core of the rush hour as headways become tighter. One surprising result is given for the F line from 9:00 to 9:29 a.m. The standard method shows little change from the previous period, but the operational method shows that service has significantly improved. Similarly, the No.4 service is shown to worsen in the 8:00 to 8:29 period, whereas the standard method shows it nearly the same as that of the previous half-hour.

Comparing lines from an operational perspective is different from the rider's view. Riders are interested in the probability of on-time arrival; the No.1 provides much better service than the No. 5. However, in operational terms, lines can be compared only in the context of their infrastructures and service configurations. The performance of two lines can be compared only after controlling for exogenous factors like the number of merges, distance, headways, and equipment reliability. This kind of analysis can only be performed with a sophisticated causal model and a large data base.

Total-Trip OTP

This version of OTP provides the crucial link between travel times and the regularity of service (platform waiting times).

TABLE 2 OPERATIONAL OTP OF THE F AND NO.4 LINES

Scheduled Arrival at CBD Location (a.m.)	Standard OTP (%)		Operational OTP (%)	
	F North[a]	No.4 South[b]	F North[a]	No.4 South[b]
6:00-6:29	75.9	92.4	70.4	89.4
6:30-6:59	75.0	90.1	72.2	87.9
7:00-7:29	77.8	85.5	73.3	80.9
7:30-7:59	83.3	80.2	75.0	74.1
8:00-8:29	64.8	77.9	53.7	70.8
8:30-8:59	64.3	50.6	50.8	41.6
9:00-9:29	63.0	52.2	61.1	42.5
9:30-9:59	88.6	92.0	81.4	89.8

[a]F northbound serves Brooklyn and the Lower East Side of Manhattan.
[b]No. 4 southbound serves Upper Manhattan and the Bronx.
SOURCE: MTA–IG Analysis of 1988 NYCTA Interval Sheets.

A train is late if the actual travel time plus the actual wait exceeds the scheduled travel time plus the scheduled wait by more than 5 min. The actual wait is half the headway—the average wait of all passengers on the train, assuming a uniform arrival rate of passengers onto the platform and no riders left on the platform by the previous train.

The method of combining travel times and waiting times permits analytical treatment of useful schedule adjustments. An en route schedule adjustment may be required when dispatchers learn that a train must be cancelled. The train preceding the cancelled train is held at a station to close the gap behind it. This process delays the travel time of the passengers on the train, but spares many passengers down the line a longer waiting time. In effect, a schedule adjustment spreads out the delay over two or more trains. Schedule adjustments made at terminals may not affect travel times, but they cause some passengers to be late because of increased waiting times. By separately calculating the actual travel and the actual waiting times, the model accounts for the increased lateness for some passengers as well as the decreased wait for others because of the schedule adjustments.

Under the definition of total-trip OTP, passengers are late if

$$(ATR + AW) - (STR + SW) > 5 \text{ min}$$

where

ATR = actual travel time,
AW = average actual wait, or one-half the actual headway,
STR = scheduled travel time, and
SW = average scheduled wait, or one-half the scheduled headway.

Each cancelled train represents a passenger group for which travel and waiting times are calculated separately. The passengers who would have boarded the cancelled train are attributed the travel time of the next available train. Some cancelled trains are not late under this definition. If the next available train reached the key point 5 min or less from the scheduled time of the cancelled train, then it makes no sense in this model to record this passenger group as late.

The incorporation of waiting times into the calculation of OTP allows the model to account for extra trains sent out to fill a gap in service. The extra trains will reduce the waiting time of passengers, improving OTP under the total-trip method.

The results of this analysis are presented in Table 3. In general, the standard and total-trip methods yield similar results. The apparently larger differences in the shoulder of the rush are probably caused by larger headways.

However, passengers cannot be assumed to be distributed equally among trains. When service is not timely, it tends to be more erratic. Headways become uneven, and trains are often bunched together. The waiting time for a bunched train can be as short as 1 min. If that train's travel time was 7 min more than scheduled but the waiting time was 2 min less than scheduled, it would be counted as on time (only 5 min late) by the total-trip method. The reduced waiting time offsets the longer travel time. This effect seems to be the reason the total-trip method gives results similar to the standard method. However, bunched trains do not represent better service.

TABLE 3 STANDARD OTP VERSUS TOTAL-TRIP OTP FOR THE F AND NO.4 LINES NORTHBOUND

Scheduled Arrival at CBD Location (a.m.)	Standard OTP (%)		Total Trip OTP (%)	
	F	No.4	F	No.4
6:00-6:29	75.9	90.9	63.2	88.5
6:30-6:59	75.0	90.9	75.0	86.4
7:00-7:29	77.8	89.8	76.4	86.4
7:30-7:59	83.3	82.1	81.5	80.4
8:00-8:29	64.8	86.4	66.7	81.8
8:30-8:59	64.3	60.5	66.7	62.0
9:00-9:29	63.0	36.3	65.7	42.1
9:30-9:59	88.6	63.6	82.9	72.2

NOTE: The F and No.4 northbound lines serve Brooklyn.
SOURCE: MTA–IG Analysis of 1988 NYCTA Interval Sheets

Therefore, a method must be devised to account for the effects of irregular service on passenger OTP. One way to account for this effect is to weight individual trains by the number of passengers estimated to be on a train; such an approach is outlined in the following section.

Weighted Total-Trip OTP

Assuming a uniform rate for passenger entries into the station, a longer wait will cause passengers to accumulate on station platforms. As a result, a train that comes after a delay will be more crowded than the trains following. Trains with longer headways will be more crowded than trains bunched behind. For example, if 20 passengers arrive every minute, a train with a headway of 8 min will pick up 160 passengers; a train following with a headway of 2 min will pick up only 40 passengers. The headway of the first train is four times longer than the headway of the second; similarly, the number of passengers is four times as great.

Therefore, headways provide a simple method for weighting individual trains by their passenger loads. As long as the subject is a single route within a narrow time frame, the actual rate of passenger arrival is not needed: the ratio of headways will always give the ratio of passengers. Instead of calculating OTP by the standard formula

$$\text{OTP} = \frac{\text{number of trains on time}}{\text{total number of trains}}$$

the following is used:

$$\text{OTP} = \frac{\text{total headway of on-time trains}}{\text{total headway of all trains (late and on time)}}$$

This is the same as

$$\text{OTP} = \frac{\text{total number of passengers on time}}{\text{total number of passengers (late and on time)}}$$

An example may be helpful. If two trains are late and two are on time, the standard OTP would be 50 percent. However, if the on-time trains were bunched behind the late trains so that the headways of the on-time trains are 2 and 3 min but the headways of the late trains are 6 and 4 min, then the weighted method would give on-time performance as $(2 + 3)/(2 + 3 + 6 + 4) = 33$ percent. More passengers were on the delayed trains.

If the weighted total trip method is used to estimate OTP for the northbound No.4, the results are consistently lower than those from the simple total-trip method. They are sometimes higher and sometimes lower than those of the standard method. Between 6:30 and 7:00 a.m., the standard OTP is 90.9 percent, the total-trip OTP is 86.4 percent, and the weighted total-trip OTP is 82.2 percent. The weighted method's estimate is almost 10 percent lower than that of the standard OTP. However, between 9:00 and 9:30 a.m., the standard OTP is 36.3 percent, whereas the weighted total-trip OTP is 39.9 percent (about 10 percent higher).

The major impact of irregular service on OTP is the creation of overcrowded conditions such that passengers cannot board the next available train. When this situation occurs, the waiting times and the number of late trains are larger than either the total-trip or weighted total-trip method estimates, and OTP consequently is lower. To deal with this issue, an empirical study must be made of the relationships between the distribution of headways, the distribution of passengers, and operating capacity.

AVERAGE DELAY

Reliable figures on delays are rarely presented publicly, and when they are, they focus on the delay of trains, not passengers. Because cancelled trains never reach the destination terminal, no lateness figure is assigned to them, and they are dropped from the analysis. This method is used by the Long Island Rail Road and Metro-North, the two commuter railroads within the MTA. The NYCTA does not report the average delay for its lines or for the subway system.

The average delay of passengers is a critical variable both for purposes of evaluation of service by oversight or consumer groups and for management in appraising operational strategies to minimize inconvenience to passengers. Table 4 presents the results for two routes. The average delay can be as much as 27 percent higher with cancellations included (F line, 6:30 to 6:59). A graph of the delay distributions could be expected to look very different.

Both routes tend to have larger average delays when delays due to cancellations are included, but this result is not always

TABLE 4 AVERAGE DELAY PER LATE TRAIN

Scheduled Arrival at CBD Location (a.m.)	Ignoring Cancellations (min)		Imputing Time to Cancellations (min)	
	F North[a]	No.4 South[b]	F North	No.4 South
6:00-6:29	10.8	11.0	11.1	12.0
6:30-6:59	7.8	10.3	9.9	10.3
7:00-7:29	7.4	9.3	9.0	9.2
7:30-7:59	8.7	9.1	9.4	10.4
8:00-8:29	9.2	9.6	9.7	11.1
8:30-8:59	8.0	9.2	9.2	11.0
9:00-9:29	9.6	9.8	10.6	10.8
9:30-9:59	7.7	6.3	8.3	6.3

[a]F northbound serves Brooklyn and the Lower East side of Manhattan.
[b]No.4 southbound serves Upper Manhattan and the Bronx.
SOURCE: MTA–IG Analysis of 1988 NYCTA Interval Sheets

the case. The average delay decreased for the No.4 southbound (7:00 to 7:29), because abandoned trains resulted in delays less than the average. Of course, the average delay figures could change even more if weighted by passengers and corrected for overcrowding.

CONCLUSIONS AND DIRECTIONS

The MTA–IG model's passenger orientation is expressed in fundamental aspects of the analytical framework, such as the choice of measurement points; the definition of a trip; the use of time periods; the treatment of cancellations, bypasses, and extra service; the inclusion of waiting times; and the search for a method of weighting measures by passenger loads. The major area for further development involves research to account for the effects that overcrowding and irregular headways have on passengers.

The analytical objective of measuring the passenger's experience initiated a search for ways to weight the train data, to translate the probabilities for trains' being on time into the probabilities that passengers were on time. A weighting procedure using aggregate passenger volume and aggregate OTP data was found to be inadequate for measuring the probability of on-time arrivals for passengers. Such a statistic will add little information to the standard method, which already accounts in part for differences in passenger volume by graduations in service frequency. For perfectly regular service with no overcrowding, a weighting system would be unnecessary. However, overcrowding and irregular headways, with a resulting uneven passenger distribution, make the development of a weighting method necessary. A weighted total-trip method was developed to address uneven passenger loads. This method is a good first step, but it needs empirical verification before application.

The relationship between headway variance and passenger load distribution should be studied to determine how crowded a train will be, given its own headway and the headway of trains preceding it. This task and exploring the issue of passenger loads under constraints of operating capacity (i.e., the problem of the first or subsequent trains being too crowded for some riders to board) represent the main directions for further study.

The actual passenger arrival rates on a given day cannot be known. Assumptions must be made on the basis of the most recent traffic-checking data. A passenger-oriented model, therefore, requires an adequate traffic-checking program. For the most comprehensive system, precise, up-to-date estimates are needed for every line in every time period. However, considerable analytical power can be realized with knowledge of relative passenger volumes (i.e., knowing the ratios of passengers from one time period to another and line-to-line ratios). This approach is adequate for most purposes and does not require such an intensive traffic checking effort. Ridership growth on the system may not change the distribution in relative terms.

The model's passenger orientation does not preclude measuring service in operational terms. Indeed, the data base and analytical framework needed to support the model permit impressive flexibility in analyzing service. A number of different methods of calculating OTP were examined in this paper. No single method is recommended for operating agencies. Some methods are more suited to particular diagnostic operational uses, but certainly the passenger orientation is critical for questions of operational effectiveness, especially in the context of a strategic management approach.

One of the most valuable aspects of the sample-based, analytical approach used to support the MTA–IG model is its flexibility. Operating agencies usually collect performance data according to written standard procedures. For example, NYCTA field personnel phone in how many trains were late or cancelled within a certain time period, a simple and efficient (though often inaccurate) procedure that, unfortunately, predetermines the scope of analysis. The statistical data base used here permits a wide range of analyses using different definitions of OTP, different time periods, and combinations of merging services. The analytical applications go beyond OTP; headways, waiting times, travel times, cancellations, delay recovery, schedule adequacy, and other statistics can all be measured. The MTA–IG is also preparing reports on service regularity and passenger waiting times and developing the data base as a comprehensive causal model.

The performance measurement system described here is more complicated to construct at present than the standard procedures used by most operating agencies. Even though it is based on a sample, considerable effort is required to build the data base and maintain its accuracy. The production of results requires a certain time lag, while operating agencies often need timely feedback. Why should a transit agency invest time and money in such a system?

Operating agencies stand to benefit the most from such an analytical system. Because reliable passenger service is their primary mission, at least in theory, measuring the passenger's experience should be an important goal.

Expressing performance in these terms also more accurately quantifies the expected benefits and their link to specific capital investments. Proposed performance benefits to riders can be instrumental in persuading UMTA and oversight agencies of the importance of proposed projects. A measurement system that overestimates performance by discounting the effect of irregular service on passengers or levels out the variation in performance by too much aggregation is also likely to underestimate the benefits to passengers of new policies or capital improvements. Operating agencies can also benefit from more sophisticated and flexible analytical methods for operations analysis, scheduling, and planning.

The potential users of such a system are not limited to operating agencies. In the contemporary institutional landscape, a number of federal, state, and local agencies, like the MTA–IG, have oversight responsibilities for monitoring program expenditures and program outcomes. Also, citizen activist organizations independently monitor service. For these organizations, the essential requirements are accuracy and an evaluation expressed in terms of the service experienced by passengers. It would be sufficiently timely to produce statistical results for a given year early in the following year. The MTA–IG's experience suggests that this procedure is entirely feasible after an initial period of organizational development and research.

The most problematic issue for operating agencies is to meet their needs for reporting on performance within 24 hr. This responsibility can be accomplished for the proposed system only through technological developments—the selective use of accurate automated vehicle monitoring systems and computer-supported dispatching at terminals and selected intermediate stations. These long-term goals require capital investment and employee development. A transit agency planning to introduce these technologies should consider the analytical capabilities they provide to improve the agency's understanding of service delivery problems and to evaluate managerial strategies.

DISCUSSION

P. TAKIS SALPEAS AND VUKAN R. VUCHIC
FERP, Southeastern Pennsylvania Transportation Authority, 5800 Bustleton Ave., Philadelphia, Pa. 19149; Department of Systems, 113 Towne Bldg., University of Pennsylvania, Philadelphia, Pa. 19104-6315.

The authors correctly state that the definition of reliability in transit system evaluation requires additional attention. The purpose of this discussion is to contribute some suggestions toward further development of the concept of transit service reliability.

Evaluation of transit service quality with respect to a performance characteristic can be classified into two different dimensions. First, that performance characteristic has to be defined and measured by itself; second, it can be weighted by its impact on passengers. The latter aspect involves volumes and characteristics of passengers affected by the service performance.

The authors point out the problem of finding the appropriate location for measuring reliability, which is defined as the percentage of trains arriving within 0 to 4 min from the scheduled time. They correctly suggest that the on-time performance should be measured at the point where most passenger trips terminate, rather than at the end of the line, which may be in the suburbs where passenger volume is extremely low. In addition, reporting should be done for shorter intervals (i.e., 30 min rather than 1 hr).

The importance of reliability depends partly on service headway. With short headways, delays that approach the headway in length are not felt very much because passengers, except the ones on the delayed vehicle, may not notice that difference. The vehicle they take may not be the scheduled one, but it serves them close to schedule. Under such conditions, maintaining regularity of service (uniform headways) becomes more important than maintaining schedules. On lines with long headways, however, reliability is of utmost importance; passengers rely on that service and often have no alternative.

Passengers sometimes perceive the impact of a delay according to the duration of their trip. A 10-min delay on a 20-min trip may be more irritating than the same delay on a 60-min trip. Yet, reliability of service is equally important for all trip lengths, because the passenger is equally concerned with arriving on time regardless of the distance traveled.

On the passenger side, passenger volume expresses the breadth of the impact of service reliability, as the authors correctly point out, whereas the type of impact can be measured by the sensitivity of passengers to reliability (or lack of it). Sensitivity is a function of the consequences of low reliability. If the consequences are serious, passengers are very irritated by any delays. An extreme example is travel to the airport, where delays on transit lines may cause the passenger to miss a flight. The passenger traveling for leisure or casual shopping is much less sensitive to a similar delay.

Passenger characteristics that influence this sensitivity include such factors as trip purpose, trip duration, and ridership composition. Trip purposes could be classified and greater weight be given to work, business, and school trips than to social and shopping trips. The second characteristic, ridership composition, can be included through grouping by age; for senior citizens, on the average, travel reliability is less important than for persons in working and school ages.

The importance, complexity, and multiple interrelationships of the influencing factors suggest that the reliability of each line should be measured by models that include the most relevant of these factors. However, this can be impractical because of extreme complexity of the required data collection and analyses.

The model can be simplified by use of fewer factors that could act as proxies for all the discussed elements. Further research should be done to derive these elements.

For example, it would be impractical to try to measure the percentages of passengers by trip purposes, passenger age, and other characteristics in measuring impacts of reliability. However, it may be practical to distinguish reliability during peak hours, dominated by work and school travel, from reliability of service during offpeak hours, used more by discretionary travelers.

It is interesting that the authors' much more sophisticated method for computing reliability has not resulted in very different findings from those obtained through conventional reliability measures. That may indicate that the conventional methods are robust enough to produce reasonable results. Yet, the increasing need for more sophisticated analyses of reliability requires further efforts to develop more complex and sensitive, yet practical, methods.

The authors of this discussion are solely responsible for its contents and conclusions, which may not represent the official view or policies of SEPTA.

AUTHORS' CLOSURE

The discussants have provided interesting extensions of the conceptual approach presented in the paper. The issue of the passenger's sensitivity to poor reliability opens another fruitful area for analysis. Although the paper's use of demographic characteristics is limited to passenger distributions and volumes, the concept of passenger sensitivity brings into the analysis the realization that passengers in different circumstances will respond differently to delays. This line of thought also suggests that the relationship between the passenger's tolerance of delays and the magnitude of the delay may not be linear. Passengers may be inconvenienced but tolerant of small delays, but increasingly dissatisfied at higher levels. An OTP measure that acknowledges such subjective factors might require giving greater weight to larger delays.

The discussants identify three factors that influence this sensitivity, or tolerance of delay—trip purpose, trip duration, and ridership composition. The authors suggest that trip purpose could be included by giving greater weight "to work, business, and school trips than to social and shopping trips." The analytical problems are to find a good proxy for these different purposes and to decide how much to weight them. A possible solution might be to use a different standard (i.e., how many minutes is late?) for reliability for midday and offpeak than for rush hours, factoring in estimates of ridership composition. The variation in standards could be derived from a survey of passengers at different times.

In considering trip duration, many passengers use linked trips. Conventional measures focus on single lines, but transfers introduce a new element. Using linked trips can even lead to a more adequate appraisal of service experienced by passengers using buses as feeders to the subway.

The discussants also note the robustness of the measures produced by conventional methods. Our efforts to combine waiting times with travel time—the total-trip method—had little effect on our estimates of reliability in the core of the rush hour. Larger differences were produced for the shoulders of the rush. In subsequent analysis, we found up to 14 percent differences for given time periods on certain lines. This variation suggests that the total-trip method is most useful when headways are larger, especially during midday and evenings. The differences between the total-trip and conventional methods are greater still when a weighting method is used, as discussed in the paper.

For measuring reliability in the core of the rush hour, our more sophisticated methods currently yield results not much different from the conventional method. However, in the core of the rush, overcrowding causes passengers lateness that cannot be detected by merely timing the trains. Considerable numbers of passengers cannot board the first train that passes because of overcrowding. Therefore, we believe it is necessary to increase the sophistication of our method rather than rely on the conventional approach, because the latter fails to account for the full extent of delays experienced by passengers.

Publication of this paper sponsored by Committee on Rail Transit Systems.